



UNIVERSIDADE FEDERAL DO PARANÁ

PEDRO BEBER DE QUEIROZ VIDAL

A COMPARATIVE STUDY ON SYNTHETIC FACIAL DATA GENERATION TECHNIQUES
FOR FACE RECOGNITION

CURITIBA PR

2025

PEDRO BEBER DE QUEIROZ VIDAL

A COMPARATIVE STUDY ON SYNTHETIC FACIAL DATA GENERATION TECHNIQUES
FOR FACE RECOGNITION

Texto final referente à disciplina de TCC2.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. David Menotti.

CURITIBA PR

2025

Universidade Federal do Paraná
Setor de Ciências Exatas
Curso de Ciência da Computação

Ata de Apresentação de Trabalho de Graduação II

Título do Trabalho: **A COMPARATIVE STUDY ON SYNTHETIC FACIAL DATA GENERATION TECHNIQUES FOR FACE RECOGNITION**

Autor(es):

GRR 20205105 Nome: PEDRO BEBER DE QUEIROZ VIDAL

GRR _____ Nome: _____

GRR _____ Nome: _____

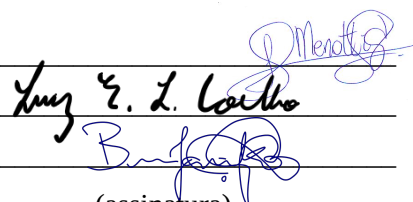
Apresentação: Data: 05 / 02 / 2025 Hora: 09:00 Local: <https://meet.google.com/wea-aawa-cyq>

Orientador: DAVID MENOTTI GOMES

Membro 1: LUIZ EDUARDO LIMA COELHO

Membro 2: BERNARDO JANKO GONÇALVES BIESSECK

(nome)


(assinatura)

AVALIAÇÃO – Produto escrito	ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo (00-40)				35
Referência Bibliográfica (00-10)				10
Formato (00-05)				05
AVALIAÇÃO – Apresentação Oral				
Domínio do Assunto (00-15)				15
Desenvolvimento do Assunto (00-05)				05
Técnica de Apresentação (00-03)				03
Uso do Tempo (00-02)				02
AVALIAÇÃO – Desenvolvimento				
Nota do Orientador (00-20)		*****	*****	20
NOTA FINAL	*****	*****	*****	95

Pesos indicados são sugestões.

Conforme decisão do colegiado do curso de Ciência da Computação, a entrega dos documentos comprobatório de trabalho de graduação 2 deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação: Trabalho Conclusão de Curso; informar os interessados: nome do aluno e o nome do orientador; anexar esta ata escaneada e a versão final do pdf da monografia do aluno.; Tramita o processo para CCOMP (Coordenação Ciência da Computação).

ACKNOWLEDGEMENTS

Gostaria de expressar minha profunda gratidão aos meus pais, cujo apoio incondicional tem sido fundamental em todas as etapas da minha vida. Suas palavras de encorajamento e sua presença constante foram essenciais para que eu pudesse chegar até aqui. Vocês são minha inspiração e meu sustentáculo.

Eu também gostaria de agradecer ao Professor David Menotti, cuja orientação durante a escrita deste trabalho foi de valor inestimável. Seu conhecimento, paciência e entusiasmo não apenas me ajudaram a superar desafios, mas também a apreciar ainda mais o processo de pesquisa. Obrigado por acreditar em mim e por fornecer uma orientação tão valiosa.

RESUMO

O reconhecimento facial se tornou um método amplamente utilizado para autenticação e identificação de usuários, com aplicações em vários domínios, como acesso seguro e localização de pessoas desaparecidas. O sucesso dessa tecnologia é amplamente atribuído ao aprendizado profundo, que aproveita grandes conjuntos de dados e funções de perda eficientes para obter recursos discriminativos. Apesar de seus avanços, o reconhecimento facial ainda enfrenta desafios em áreas como explicabilidade, viés demográfico, privacidade e robustez contra envelhecimento, variações de pose, mudanças de iluminação, oclusões e expressões. Além disso, o surgimento de regulamentações de privacidade levou à depreciação de vários conjuntos de dados bem estabelecidos, levantando preocupações legais, éticas e de privacidade. Para abordar essas questões, a geração de dados faciais sintéticos foi proposta como uma solução. Essa técnica não apenas atenua as preocupações com a privacidade, mas também permite uma ampla experimentação com atributos faciais tendenciosos (i.e. tom de pele e cabelo), ajuda a aliviar o viés demográfico e fornece dados suplementares para melhorar os modelos treinados em dados reais. Essas características foram consideradas em competições, como o *Face Recognition Challenge in the Era of Synthetic Data (FRCSyn)* e o *Synthetic Data for Face Recognition Competition (SDFR)*, que foram organizadas para explorar as limitações e o potencial da tecnologia de reconhecimento facial treinada com dados sintéticos. Este estudo compara a eficácia de conjuntos de dados faciais sintéticos estabelecidos com diferentes técnicas de geração em tarefas de reconhecimento facial. Foram avaliadas as métricas de precisão, rank-1, rank-5 e taxa de verdadeiro positivo (TPR) a uma taxa de falso positivo (FPR) de 0,01%, em oito conjuntos de dados principais, fornecendo uma comparação de abordagens que não são explicitamente contrastadas na literatura. Os experimentos destacam as várias técnicas usadas para abordar o problema da geração de dados faciais sintéticos e apresentam uma avaliação abrangente do campo. Os resultados demonstram a eficácia de vários métodos na geração de dados faciais sintéticos com variações realistas, destacando as várias técnicas usadas para abordar o problema. Concluiu-se que as técnicas de geração de dados sintéticos, como modelos de difusão, GANs e modelos 3D, avançaram na replicação da complexidade do mundo real para reconhecimento facial. No entanto, a lacuna em relação aos dados reais ainda existe, exigindo pesquisas futuras.

Palavras-chave: Reconhecimento facial. Biometria. Dado Facial Sintético.

ABSTRACT

Facial recognition has become a widely used method for user authentication and identification, with applications in various domains such as secure access and missing person location. The success of this technology is largely attributed to deep learning, which leverages large datasets and efficient loss functions to achieve discriminative features. Despite its advances, facial recognition still faces challenges in areas such as explainability, demographic bias, privacy, and robustness against aging, pose variations, lighting changes, occlusions, and expressions. Furthermore, the emergence of privacy regulations has led to the deprecation of several well-established datasets, raising legal, ethical, and privacy concerns. To address these issues, synthetic facial data generation has been proposed as a solution. This technique not only mitigates privacy concerns but also allows for extensive experimentation with biasing facial attributes (i.e. skin tone and facial hair), helps alleviate demographic bias, and provides supplementary data to improve models trained on real data. This characteristics were considered in competitions, such as the Face Recognition Challenge in the Era of Synthetic Data (FRCSyn) and the Synthetic Data for Face Recognition Competition (SDFR), have been organized to explore the limitations and potential of facial recognition technology trained with synthetic data. This study compares the effectiveness of established synthetic facial datasets with different generation techniques on facial recognition tasks. Were evaluated the accuracy metric, rank-1, rank-5, and True positive rate (TPR) at a False positive rate (FPR) of 0.01%, on eight leading datasets, providing a comparison of approaches that are not explicitly contrasted in the literature. The experiments highlight the various techniques used to address the problem of synthetic facial data generation and present a comprehensive assessment of the field. The results demonstrate the effectiveness of various methods in generating synthetic facial data with realistic variations, highlighting the various techniques used to address the problem. It was concluded that synthetic data generation techniques, such as diffusion models, GANs, and 3D models, have advanced in replicating real-world complexity for facial recognition. However, the gap to real data still exists, requiring future research.

Keywords: Facial recognition. Biometrics. Synthetic Facial Data.

LIST OF FIGURES

2.1	An example of a CNN architecture, proposed by LeCun et al. (1998)	16
2.2	Overview of a facial recognition system. From Guo and Zhang (2019).	18
3.1	The sampling approach starts with a given off the shelf generator, that outputs a face image X_{id} , and a style image X_{sty} . The second diffusion model receive those images and combine the identity from X_{id} and the style from X_{sty} . By repeating this process multiple times, a labeled synthetic face dataset can be created. From Kim et al. (2023)	20
3.2	The proposed SynFace framework begins by integrating identity mixup into DiscoFaceGAN, resulting in the Mixup Face Generator. This generator is capable of producing face images that exhibit a variety of identities, including their intermediate states. Subsequently, these synthetic face images are combined with a limited number of real face images through a process called domain mixup, which helps to mitigate the domain gap. After sampling, the mixed face images are then fed into a feature extractor, which derives the corresponding features, that are used to compose the margin based loss. From Qiu et al. (2021).	21
3.3	Each row illustrates the same individual depicted with various accessory configurations (left figure). These accessories encompass attire, eyewear, cosmetics (such as eyeshadow and eyeliner), as well as facial and head adornments. Additionally, the color, density, and thickness of facial and head hair are randomized. The hairstyle is altered only when the chosen accessory clashes with the original style. These images (right image) demonstrate how the same face can appear markedly different based on pose, expression, lighting, background, and camera settings, thereby promoting the network’s ability to learn a robust embedding. From Bae et al. (2023)..	21
3.4	They first generate images containing limited intra class variations, using a GAN. In a second stage these images are used to train a text-conditioned diffusion model that outputs images with realistic intra class variations, that once filtered, will compose the final dataset. From Melzi et al. (2023).	22
3.5	They utilize the StyleGAN2 to generate the images. This method receives as input the identity labels, that are embedded and concatenated with the latent variables. After sampling, they conducted training using multi-class classification, knowledge transfer and combined learning. From Boutros et al. (2022).	22
3.6	A summary of the generator and discrimination training procedure. The models are trained concurrently, where the updated pipelines are marked as green boxes. The identity driven loss and batch normalization statistics are used to train the generator as an add to generator loss. After training, a synthetic face dataset is created, aligned and cropped, and feed to FR model trained with margin-penalty based loss function. From Kolf et al. (2023).	23

- 3.7 The illustration is divided into two segments. The upper segment details the training process, where a denoising U-Net is guided by contextual information derived from a pre-trained Face Recognition model's features. This training occurs within the latent domain of a pre-trained Autoencoder, with the diffusion process providing targets for the reverse learning sequence. The lower segment explains the sample creation process, where the trained Diffusion Model can generate samples using three identity contexts: real, two-stage, or synthetic uniform. By maintaining a fixed identity context and modifying the noise, diverse samples for the same identity are achievable. From Boutros et al. (2023a). . . . 24
- 3.8 For each identity a , a random vector z_a is sampled and mapped to initial latent $w_a^{(0)}$. An image i_a is generated from w_a and its embedding e_a computed. Loss functions on embedding space E and latent space W are used to update latents $w_a^{(t)} \rightarrow w_a^{(t+1)}$ over N_{iter} iterations. This method is used to generate the interclass variations (left figure). For each identity a and variation α , initialize latent $w_a^{\alpha(0)}$ near w_a^{ref} (right figure). Losses pull embeddings towards e_a^{ref} , latents towards average w_{avg} , and repel nearby latents. Repeat for all a and α over N_{disp_iter} iterations. This method is used to generate the intraclass variations. From Geissbühler et al. (2024). . . . 25
- 3.9 The images are firstly generated with a synthesizer, its embeddings are computed, and the optimization process is applied. As a result, the interclass variation is achieved and the embeddings are used to generate the final dataset images. From Shahreza and Marcel (2024). 25
- 3.10 The IM feature is computed by a facial recognition model, and then it is expanded into a feature map. The latter is processed by a feature mask autoencoder (fMAE), where the rows are randomly masked. Then they are fed to an image decoder, that reconstructs the pixels. The training is conducted calculating the MSE and cosine similarity between the reconstructed image and the ground truth, and the perceptual loss and GAN loss are used to ensure a correct facial structure and increase the sharpness of the generated images. From Wu et al. (2024). 25
- 3.11 The Arcface embedding are concatenated to the CLIP input, that outputs conditional embeddings that are used for cross-attention control. The UNet and the encoder are optimized using a million scale dataset, and further finetuned on a high resolution dataset, without any text annotations. From Papantoniou et al. (2024). 26
- 3.12 The training predictions for race (R^*), gender (G^*) and age (A) are outputs of a CLIP model. Next to ensure Facial recognition consistency, its used a FR pipeline to refine the race and gender labels, as well as computing the ID and divergence (DS) labels. These labels are used to train the diffusion models in stage 1 and 2. The inference procedure consists of using the balanced synthetic identities provided in stage 1, filtered and processed, to generated synthetic embeddings. These and randomly sampled A and DS are used as input to the second stage diffusion model to generate the synthetic dataset, that a passed to the second filtering stage to create the filtered dataset. From Yeung et al. (2024). 26

3.13	The pipeline was trained with real prototypes indicated as WR , and k prototypes for virtual IDs, denoted as WV . The virtual embedding $f'_{FR}(x_j)$ is designed to simulate the distribution of real embeddings. Subsequently, a diffusion model is used to generate the synthetic images based on the virtual prototypes. From Kim et al. (2024).	27
5.1	Achieved ROC curves on LFW, CALFW and AGEDB for the trained datasets. TPR means True Positive Rate and FPR means False Positive Rate	36
5.2	Achieved ROC curves on CFPFP and CPLFW for the trained datasets. TPR means True Positive Rate and FPR means False Positive Rate	36
5.3	Achieved ROC curves on IJBB and IJBC for the trained datasets. TPR means True Positive Rate and FPR means False Positive Rate.	36

LIST OF TABLES

3.1	Table summarizing the objectives and details of facial recognition datasets. . . .	28
4.1	Summary of facial Recognition datasets, evaluation protocols, and their challenges.	33
5.1	Performance result across selected datasets on mainstream datasets. * indicates that the results were taken from the original article itself. The mainstream dataset are called Labelled Faces in the Wild (LFW), Cross-Pose LFW (CPLFW), Celebrities in Frontal-Profile (CFP) (protocol FP), Cross-AgeLFW (CALFW), Age Database (AGEDB).	34
5.2	Performance result across selected datasets on IARPA Janus Benchmark-C (IJBC), IARPA Janus Benchmark-B (IJBB), TinyFace R1, and TinyFace R5.	35

LIST OF ACRONYMS

FR	Facial recognition
FRCSyn	Face Recognition Challenge in the Era of Synthetic Data
SDFR	Synthetic Data for Face Recognition Competition
TPR	True positive rate
FPR	False positive rate
NIST	National Institute of Standards and Technology
PCA	Principal component analysis
LDA	Linear Discriminant Analysis
SVM	Support Vector Machine
ICA	Independent component analysis
LBP	Local binary patterns
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
SGD	Stochastic gradient descent
Adam	Adaptive Moment Estimation
CMC	Cumulative Matching Characteristics
ROC	Receiver-operating characteristic curve

CONTENTS

1	INTRODUCTION	12
1.1	MOTIVATION	13
1.2	OBJECTIVE	13
1.3	OUTLINE.	14
2	THEORETICAL FUNDAMENTATION	15
2.1	FACIAL RECOGNITION WITH HANDCRAFTED FEATURES	15
2.2	CONVOLUTION NEURAL NETWORKS	15
2.3	FACIAL RECOGNITION WITH DEEP LEARNING	16
2.4	FACIAL RECOGNITION SYSTEM PIPELINE OVERVIEW.	18
2.4.1	Preprocessing step	18
2.4.2	Deep learning feature extractor and facial matcher	18
3	STUDIED WORKS	20
4	METHODOLOGY	30
4.1	IMAGE PREPROCESSING	30
4.2	TRAINING CHOICES	30
4.2.1	Used Backbone	30
4.2.2	Loss Function	30
4.2.3	Augmentation Applied	31
4.2.4	Method Description	32
4.3	EVALUATION PROTOCOL	32
5	RESULTS AND DISCUSSION.	34
6	CONCLUSION	39
	REFERENCES	40

1 INTRODUCTION

Face recognition (FR) has rapidly become a prevalent method for authenticating and identifying users due to its convenience and efficiency. This technology is widely employed in various sectors, including security, where it is used to grant access to secure facilities or devices, ensuring that only authorized individuals can enter sensitive areas or use specific equipment. In law enforcement, face recognition plays a crucial role in identifying suspects, solving crimes, and maintaining public safety by matching faces captured in surveillance footage with criminal databases. Additionally, it serves humanitarian purposes, such as locating missing persons by comparing images with those in public records or social media. The integration of face recognition technology into everyday life is expanding at an unprecedented rate, driven by advancements in artificial intelligence and machine learning. It is now embedded in smartphones, allowing users to unlock their devices with facial recognition, and in social media platforms, where it helps tag and organize photos.

Despite its widespread adoption and the technological advancements that have propelled facial recognition into mainstream use, this technology is not without significant challenges and controversies. One of the most pressing issues is its susceptibility to discriminatory effects and demographic bias, which can undermine its reliability and fairness. These biases often stem from unbalanced data sampling, where certain demographic groups are underrepresented, leading to skewed datasets that do not accurately reflect the diversity of the population. The processes of data collection and labeling can also introduce biases, as they may inadvertently favor certain demographic or gender groups over others.

Moreover, the approaches used in data preprocessing and modeling can exacerbate these biases, resulting in algorithms that perform unevenly across different demographic groups. This has been fundamented by a series of comprehensive studies conducted by the National Institute of Standards and Technology (NIST) in the United States, spanning from 2002 to 2019 (Grother et al., 2019a, 2018, 2019b). These studies revealed significant racial and gender biases in many widely used facial recognition algorithms, highlighting that these systems often misidentify individuals from minority groups at disproportionately higher rates compared to those from majority groups.

In addition to issues of bias, facial recognition technology raises substantial privacy concerns. The potential for privacy violations is considerable, as the technology can be used for mass surveillance without individuals' consent, leading to unauthorized tracking and profiling. These challenges underscore the urgent need for rigorous ethical standards, transparent practices, and robust regulatory frameworks to ensure that facial recognition technology is developed and deployed in a manner that is equitable, accurate, and respectful of individual privacy rights.

The facial recognition technology relies on image processing to extract features from faces. These features are then used as input to pattern recognition methods that can identify and match faces. Increasingly, these pattern recognition methods are based on machine learning, such as deep learning networks. Deep learning has been shown to be effective in extracting information from facial images. Trained on large data sets of facial images, deep learning models can learn to identify and extract a wide variety of features from faces, such as the shape of the face, the eyes, the nose, the mouth, and the eyebrows.

Face recognition is difficult because faces are complex and variable. The same face can look different depending on the viewing angle, lighting color and direction, and facial expressions. Additionally, faces can be occluded by hair, glasses, masks, or other objects.

Recently, the emergence of facial recognition technology has witnessed a transformative shift with the integration of synthetic data, marking a significant evolution in the field. As traditional methods for collecting facial data faced challenges such as privacy concerns and limited datasets, synthetic data has emerged as a groundbreaking solution. By generating computer-generated images that mimic real-world facial features, researchers and developers can now create vast and diverse datasets for training facial recognition algorithms.

This not only addresses the ethical concerns associated with using real people's data but also allows for a more comprehensive and representative training set. The incorporation of synthetic data in facial recognition has sparked significant interest among research institutions, fostering a dynamic environment of innovation and discovery. Researchers are increasingly focused on developing more accurate, inclusive, and privacy-aware facial recognition systems. As academic competition continues to drive progress, the synergy between facial recognition and synthetic data holds the potential to transform the landscape of biometric technology, providing robust and ethical solutions for various applications, ranging from security to academic research.

1.1 MOTIVATION

In research carried out on the state of the art in this topic, it was found that there are still many gaps in knowledge and multiple possibilities for advancement in this area of research, which justifies the need for further studies with the aim of improving this technology.

To address these gaps and possibilities for advancement, multiple competitions related to this topic were proposed. The 1st and 2nd editions of the FRCsyn competition (Melzi et al., 2024; DeAndres-Tame et al., 2024) aimed to answer the following questions: What are the limitations of FR technology trained only with synthetic data? Can synthetic data help alleviate current limitations in FR technology? For the first edition, the organizers proposed subtasks that invited the participants to use synthetic data alone and in conjunction with real data to mitigate demographic bias and bring performance improvement. In the second edition, they extended to an unconstrained number of synthetic images, maintaining the same objectives. With a related objective, SDFR competition (Shahreza et al., 2024) proposed that participants submit original solutions to generate synthetic data for performance improvement and mitigate the synthetic-to-real gap.

As the competitors are encouraged to submit models with already established synthetic datasets or generate new ones, a lot of characteristics need to be considered. For example, the used dataset needs to have sufficient intra-class variations and changes in pose, aging, expressions, occlusions, and illumination. Adding to that, the datasets need to have sufficient interclass variations, for the proposed models to generalize to new unseen data. Given these challenges, generating sufficient intra-class and interclass is an active area of research.

1.2 OBJECTIVE

This work aims to compare the accuracy, rank-1, rank-5, and True positive rate (TPR) at a False positive rate (FPR) of 0.01% of different generated synthetic facial datasets. Through these experiments, this work seek to provide a comprehensive assessment of the field, contrasting different approaches and highlighting the various techniques employed for this purpose.

1.3 OUTLINE

The following chapters are divided as follows: Chapter 2 gives information about the facial recognition task theoretical background, which includes the preprocessing techniques used and different loss functions and backbones employed. Chapter 3 presents an overview of facial recognition methods and synthetic dataset generation state-of-the-art, compared in this work. Chapter 4 describes the methodology used to compare the datasets, which includes the backbone and loss function used to train and the selected datasets for evaluation. Chapter 5 discusses the results obtained by comparing the different data generation methods used. Chapter 6 will present the conclusion of this work, by including any limitations found during the process.

2 THEORETICAL FUNDAMENTATION

2.1 FACIAL RECOGNITION WITH HANDCRAFTED FEATURES

Facial recognition is an active research topic in the computer vision area. At first, the problem was treated in non-deep learning ways. The method Eigenface, proposed by Turk and Pentland (1991), transformed an image to a 1D feature, and use Principal Component Analysis (PCA) to discriminate the low dimensional space. They adopted the euclidean distance to measure the similarity of a given query face in respect to a gallery. Fisherface, proposed by Belhumeur et al. (1997) pointed out that PCA maximizes the variance of all samples in the low dimensional space. Also, Eigenface did not take into account the class label, while Fisherface use this information and used Linear Discriminant Analysis (LDA) for dimensionality reduction, which maximized the interclass and intraclass variance ratio.

Another upgrade in performance was brought by Support Vector Machine (SVM). As SVM and feature extractors can be decoupled, many solutions have been proposed. Déniz et al. (2003) used Independent Component Analysis (ICA), used by Oja and Hyvarinen (2000) for feature extraction, and then used SVM to predict the face ID. Kong and Zhang (2011) designed fast least squares to accelerate the training process. Jianhong (2008) combined kernel PCA and least squares SVM to get a better result.

One of the most used feature extractor, Local binary patterns (LBP), was employed in many facial recognition methods with handcrafted features. Ahonen et al. (2004) adopted LBP to extract features of all regions from one face image, and use those features to make a histogram of features and obtain a final embedding. Wolf et al. (2008) proposed an improved descriptor based on LBP, and achieved a result with better accuracy. Tan and Triggs (2010) refined LBP and proposed Local Ternary Patterns, with a higher tolerance to image noise.

2.2 CONVOLUTION NEURAL NETWORKS

In the early 2010s, facial recognition shifted the focus to deep neural networks as the main solution to the problem. This is because the field has taken a twist with the rise of the Artificial Neural Network (ANN). This biologic inspired model has shown the potential to supplant other machine learning methods in common tasks, such as image classification and object recognition.

One of the most impressive forms of ANN architecture is the Convolutional Neural Network (CNN). These models are used to solve difficult pattern recognition tasks, with a precise and simple architecture, as showed in Figure 2.1, and are briefly described in the following paragraphs.

The convolutional layer is the primary building block of a CNN. The layer's parameters consist of a set of learnable filters (or kernels), which have a small receptive field but extend through the full depth of the input volume. As the filter slides (or convolves) around the input image, it produces a feature map that gives the responses of that filter at every spatial position. The convolution operation captures the local dependencies in the input. After each convolution operation, an activation function is applied to introduce non-linear properties to the system, enabling it to learn more complex functions. A commonly used activation function is the Rectified Linear Unit (ReLU), which applies a non-linear thresholding operation, setting all negative values in the feature map to zero while keeping the positive values.

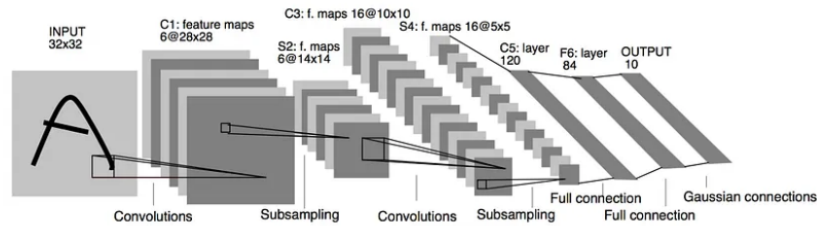


Figure 2.1: An example of a CNN architecture, proposed by LeCun et al. (1998)

The max pooling layer is responsible for reducing the dimensionality of each feature map but retains the most important information. Pooling can be of different types: max pooling, average pooling, etc. Max pooling takes the maximum value from each window of a predefined size and stride, effectively reducing the spatial dimensions of the input volume.

After several convolutional and pooling layers, the high-level reasoning in the neural network is done via fully connected layers. Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular neural networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset. In classification tasks, the final fully connected layer is often followed by a softmax (for multi-class classification) or a logistic (for binary classification) layer, which provides the classification scores for each class.

This networks need training to achieve the desired output convergence. The training process starts with a batch of images represented by a matrix of pixel values and passing through the convolutional and pooling layers, the network learns to identify various features in the input. Early layers may detect simple features like edges and curves, while deeper layers can identify more complex features like shapes or specific objects.

After feature extraction, the fully connected layers act as a classifier on top of these features and assign a probability for the input image being in a specific class. During training, the CNN uses backpropagation to adjust its weights and biases to minimize the difference between the predicted output and the actual label of the input image. This process requires a loss function to quantify the error and an optimization algorithm, like Stochastic gradient descent (SGD), Adaptive Moment Estimation (Adam) etc., to adjust the parameters based on gradients computed during backpropagation.

2.3 FACIAL RECOGNITION WITH DEEP LEARNING

Facial recognition technology has experienced remarkable advancements over the years, primarily fueled by the evolution of deep learning architectures. A significant breakthrough in this domain occurred with the introduction of AlexNet, proposed by Krizhevsky et al. (2012). AlexNet's success in achieving unprecedented results on the ImageNet dataset marked a turning point, sparking widespread adoption of deep learning methods in facial recognition and setting the stage for further innovations.

Following AlexNet, several key deep learning architectures emerged, each contributing uniquely to the field. VGGNet, developed by Simonyan and Zisserman (2014), emphasized simplicity and depth. By employing smaller 3x3 convolutional filters and deeper networks, VGGNet demonstrated that increasing network depth could enhance performance, influencing the design of future CNN's.

GoogleNet, designed by Szegedy et al. (2015), also known as Inception, introduced a novel approach with its Inception module. This architecture allowed for more efficient computation by utilizing multiple filter sizes in parallel, reducing computational costs while maintaining high accuracy. GoogleNet's efficiency made it particularly suitable for large-scale applications, further advancing the capabilities of facial recognition systems.

ResNet, introduced by He et al. (2016), addressed the vanishing gradient problem that often plagued deep networks. By incorporating residual connections, ResNet enabled the training of extremely deep networks, significantly improving accuracy across various tasks, including facial recognition. This innovation underscored the potential of deep learning architectures to push the boundaries of what was previously achievable.

In addition to these architectural advancements, notable contributions specifically targeting facial recognition emerged. DeepFace, proposed by Taigman et al. (2014), utilized a nine-layer convolutional network with a crucial facial alignment step. This approach highlighted the importance of pre-processing steps like alignment in enhancing recognition accuracy, demonstrating the potential of deep learning in refining facial recognition systems.

Another significant contribution came from the authors Schroff et al. (2015), with the introduction of FaceNet, which employed a convolutional network trained with triplet loss. This innovative loss function aimed to minimize the distance between an anchor sample and a positive sample (of the same class) while maximizing the distance between the anchor and a negative sample (of a different class), thereby enhancing the distinguishability of learned features.

Initially, facial recognition models relied heavily on the softmax loss function, combined with well-designed CNNs and large-scale training datasets. However, as the field progressed, new loss functions were developed to address challenges such as intra-class variations caused by occlusions, illumination changes, pose differences, and expression variations. A pivotal advancement came with the introduction of triplet loss in FaceNet, which provided a robust solution to these challenges by increasing the distance between positive and negative samples using a margin factor, thereby enhancing the distinguishability of the learned features, as defined in equation 2.1.

$$\|f(x_a) - f(x_p)\|_2^2 + \alpha < \|f(x_a) - f(x_n)\|_2^2 \quad (2.1)$$

The components of equation 2.1, x_a , x_p and x_n , represent the anchor, positive and negative samples, respectively, α is a margin and $f(\cdot)$ represents a nonlinear transformation embedding an image into a feature space.

Nowadays, the most prominent base loss functions are based on softmax, ie., ShephereFace, Cosface and Arcface. The ShephereFace loss function, proposed by Liu et al. (2017) introduced the idea of angular margin. Unlike traditional softmax loss, SphereFace modifies the decision boundary to include an angular margin. This modification encourages the network to learn features that are not only separable but also have a larger angular distance between them. However, this loss function required a series of approximations in order to be computed, which resulted in an unstable training of the network. Also, the authors needed to include the standard softmax loss to stabilize the training. Empirically, the softmax loss tends to dominate the training process due to the nature of the integer-based multiplicative angular margin. This margin causes the target logit curve to become extremely steep, which in turn makes it difficult for the model to converge effectively.

On the other hand, Cosface, introduced by Wang et al. (2018), directly adds cosine margin penalty to the target logit, and obtain better performance. Also, admits much easier implementation and relieves the need for joint supervision from the softmax loss. Lastly, Arcface, developed by Deng et al. (2019a), is one of the most used margin based loss function, based on

methods submitted to competitions like (Deng et al., 2021). This loss function directly optimizes the geodesic distance margin on a hypersphere. This approach ensures that the learned features are not only separable but also tightly clustered around their respective class centers, which is crucial for distinguishing between different identities.

The equation 2.2 define the mentioned margin based loss functions.

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(m_1\theta_{y_i}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{y_i}+m_2)-m_3)} + \sum_{j=1, j \neq y_i}^n e^{s \cos \theta_j}} \quad (2.2)$$

By adjusting the parameters m_1 , m_2 , and m_3 , it is possible to define different loss functions used in facial recognition:

- SphereFace: Defined by $m_1 = 1.35$, $m_2 = 0$, $m_3 = 0$
- CosFace: Defined by $m_1 = 1$, $m_2 = 0$, $m_3 = 0.35$
- ArcFace: Defined by $m_1 = 1$, $m_2 = 0.5$, $m_3 = 0$

These configurations adjust the angular margin and the loss function to enhance class discrimination in facial recognition tasks.

2.4 FACIAL RECOGNITION SYSTEM PIPELINE OVERVIEW

2.4.1 Preprocessing step

The basic steps of using deep learning to authenticate a facial image against a gallery of known faces are as shown in Figure 2.2.

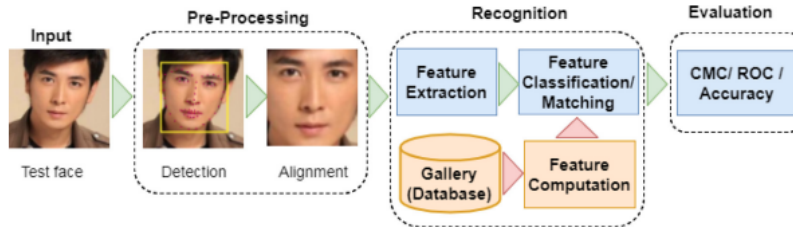


Figure 2.2: Overview of a facial recognition system. From Guo and Zhang (2019).

It starts with a data set (gallery) that includes images of known people, such as employees, students, or customers. The images should be taken in a variety of lighting conditions and from different angles. The pipeline involves preprocess the data including normalizing the images and cropping out any unnecessary background. It may also involve resizing the images and aligning them.

2.4.2 Deep learning feature extractor and facial matcher

After this, the preprocessed images are fed into a CNN for extracting features from the facial image. Regardless of the actual architecture used, all CNNs trade off spatial dimensions for channel depth through the layers. At the top-most layer, the feature maps may be flattened to form a feature vector. To authenticate a facial image, it compares the feature vector extracted from the query image to the feature vectors of the known faces in the gallery. The closest match is the identity of the person in the facial image.

Comparing the feature vectors can be done using a similarity or dissimilarity metric, such as the Euclidean distance or the cosine similarity. The most commonly used is the cosine similarity, which is a measure of the cosine of the angle between two vectors. In the case of face recognition, the cosine similarity s between two feature vectors x and y is described in Equation 2.3.

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \quad (2.3)$$

This is a measure of the similarity between the two vectors. The closer the cosine similarity is to 1, the more similar the two feature vectors are. If the objective is to find the closest identity to a given query image in a given database, this process is called facial identification, and can be evaluated using rank- x accuracy and Cumulative Matching Characteristics curves (CMC). On the other hand, facial verification involves "verifying if the person belongs to who they claim to be", and is a one-to-one matching problem, and can be evaluated using accuracy and a Receiver-operating characteristic curve (ROC).

Facial recognition technology has undergone a remarkable transformation with the advent of deep learning architectures, which have effectively surpassed earlier methods reliant on handcrafted features. CNNs have spearheaded this revolution by adeptly capturing complex patterns and subtle nuances in facial features, leading to ground-breaking innovations such as AlexNet and ResNet. These architectures have successfully addressed challenges like the vanishing gradient problem, significantly enhancing accuracy and efficiency in facial recognition systems.

The section 3, called Studied Works, describes the methods that will be compared in this study, reflecting the state of the art in synthetic facial generation techniques.

3 STUDIED WORKS

Advances related to the topic of facial recognition has continued to push the state of the art on facial recognition. Due to the release of large scales datasets of public identities crawled from the web like MS1MV2 (Deng et al., 2019a), Glint360k (An et al., 2021), Webface260M (Zhu et al., 2021) and its subset versions are the mostly commonly used for training. For testing, smaller datasets are used such as AGEDB (Moschoglou et al., 2017), CFP (Sengupta et al., 2016), LFW (Huang et al., 2008), IJB-A, IJB-C (Klare et al., 2015), IJB-B (Whitelam et al., 2017).

Also, the development of new loss functions, where the most popular are named Arcface (Deng et al., 2019a), Adaface (Kim et al., 2022), MegaFace (Meng et al., 2021), and CosFace (Wang et al., 2018) increase the accuracy percentage on the cited mainstream datasets. Also, the utilization of larger backbone deep neural networks with increased layers and parameters has facilitated the extraction of more complex features from the data, as the proposed by He et al. (2016) and Sharir et al. (2021).

However, there are problems faced with its application. Challenges related to variations in facial images, such as pose, aging, expressions and occlusions are a common problem. With the onset of Deep Learning, other challenges like the need for large amount of data, label noise presented in web-crawled datasets, and demographic disparity regarding gender and racial concerns arises. Other emerging problem are the need for informed consent, required by the legislation EU-GDPR (Voigt and Von dem Bussche, 2017). According to Boutros et al. (2023b) many of these datasets, e.g., MS-Celeb1M and VGGFace2, are retracted due to credible privacy and ethical concerns.

This movement gives growing importance for the facial recognition pipeline trained with synthetic dataset. With the development of Generative adversarial networks, 3D rendering models and diffusion models, there is an increasing interest in developing facial recognition systems trained with synthetic data, by using one of these techniques. Kim et al. (2023) proposes a diffusion model pipeline, called Dual Condition Face Generator DCFace. This pipeline consists of two stages, (i) a sampling stage generating a new identity image, and (ii) a mixing stage combining the generated image with a style image from a style bank, creating a final image that blends both style and identity information. Using this approach, the authors significantly reduced the synthetic to real domain gap. The method is described in figure 3.1.

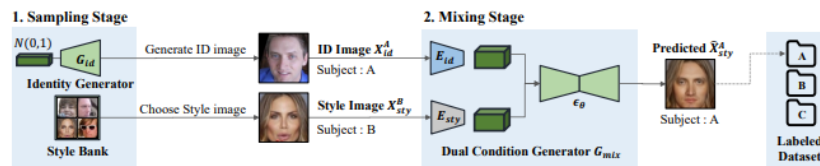


Figure 3.1: The sampling approach starts with a given off the shelf generator, that outputs a face image X_{id} , and a style image X_{sty} . The second diffusion model receive those images and combine the identity from X_{id} and the style from X_{sty} . By repeating this process multiple times, a labeled synthetic face dataset can be created. From Kim et al. (2023)

On the other hand, Synface (Qiu et al., 2021) explores the performance gap between models trained on synthetic and real face images, identifying poor intra-class variations and domain gaps as key factors. To address these, the authors introduce identity mixup (IM) and domain mixup (DM) techniques, conducting sampling with a controllable face synthesis model,

that can easily manage different factors of synthetic face generation, including pose, expression, illumination, the number of identities, and samples per identity, demonstrating significant improvements in face recognition with synthetic data. The pipeline is displayed in Figure 3.2.

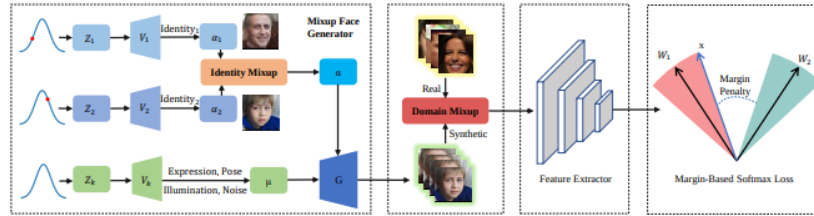


Figure 3.2: The proposed SynFace framework begins by integrating identity mixup into DiscoFaceGAN, resulting in the Mixup Face Generator. This generator is capable of producing face images that exhibit a variety of identities, including their intermediate states. Subsequently, these synthetic face images are combined with a limited number of real face images through a process called domain mixup, which helps to mitigate the domain gap. After sampling, the mixed face images are then fed into a feature extractor, which derives the corresponding features, that are used to compose the margin based loss. From Qiu et al. (2021).

Bae et al. (2023) proposed a state-of-the-art approach that uses 3D model with the objective of synthesizing high fidelity facial images. By using a computer graphics pipeline and heavy image augmentation, the authors significantly reduce the synthetic to real data accuracy gap. They also fine-tune the network to with real images obtained with consent, and achieved similar performance to models trained with real data only. The data augmentation used is described in Figure 3.3.



Figure 3.3: Each row illustrates the same individual depicted with various accessory configurations (left figure). These accessories encompass attire, eyewear, cosmetics (such as eyeshadow and eyeliner), as well as facial and head adornments. Additionally, the color, density, and thickness of facial and head hair are randomized. The hairstyle is altered only when the chosen accessory clashes with the original style. These images (right image) demonstrate how the same face can appear markedly different based on pose, expression, lighting, background, and camera settings, thereby promoting the network’s ability to learn a robust embedding. From Bae et al. (2023).

Another dataset, called GANDiffFace (Melzi et al., 2023) used a StyleGAN3 (Karras et al., 2021) to generate synthetic images and further used transformation in the latent space to control the generated attributes that serve as input to a diffusion model. This combination generates faces with desired properties, like human face realism, controllable demographic distributions, and realistic intra class variations. The Figure 3.4 illustrates the proposed pipeline.

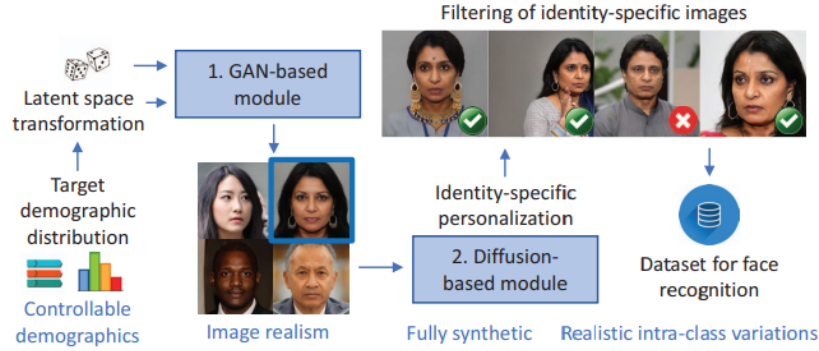


Figure 3.4: They first generate images containing limited intra class variations, using a GAN. In a second stage these images are used to train a text-conditioned diffusion model that outputs images with realistic intra class variations, that once filtered, will compose the final dataset. From Melzi et al. (2023).

The SFace, proposed by Boutros et al. (2022), is a synthetic dataset elaborated by class conditional generative adversarial network, named StyleGAN2-ADA, developed by (Karras et al., 2020). They used this dataset to train a facial recognition network under different settings: multi-class classification, label-free knowledge transfer, and combined learning of multi-class classification and knowledge transfer. The illustrative image is disposed in Figure 3.5.

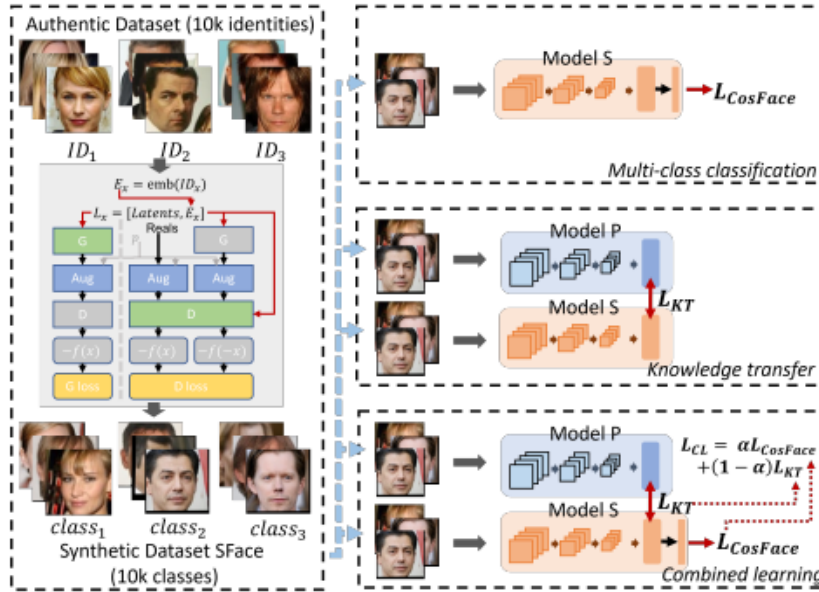


Figure 3.5: They utilize the StyleGAN2 to generate the images. This method receives as input the identity labels, that are embedded and concatenated with the latent variables. After sampling, they conducted training using multi-class classification, knowledge transfer and combined learning. From Boutros et al. (2022).

Aiming to improve the previously mentioned work, the IDnet dataset, elaborated by Kolf et al. (2023), was generated using a class-conditioned StyleGAN2-ADA. They integrate the GAN min-max game with an identity separable loss, named ID3, and a domain adaptation loss. This approach enables the generator to learn encoded identity information, generating identity-separable synthetic samples, and minimizing the domain gap between synthetic and real data distributions. The pipeline is described in Figure 3.6.

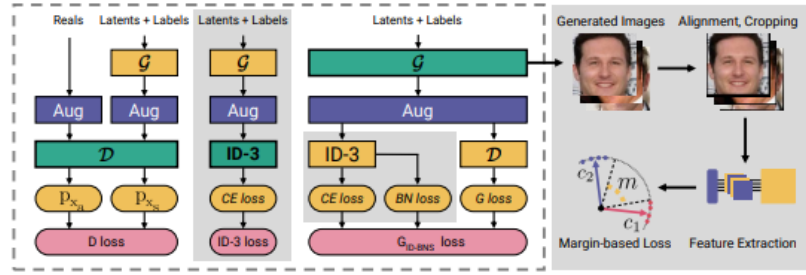


Figure 3.6: A summary of the generator and discrimination training procedure. The models are trained concurrently, where the updated pipelines are marked as green boxes. The identity driven loss and batch normalization statistics are used to train the generator as an add to generator loss. After training, a synthetic face dataset is created, aligned and cropped, and feed to FR model trained with margin-penalty based loss function. From Kolf et al. (2023).

Another work, proposed by (Boutros et al., 2023a), utilizes a diffusion model trained in the latent space of a pre-trained autoencoder. The diffusion model is also conditioned on identity context through feature representations obtained from a pre-trained face recognition model. The authors introduce a cross-attention mechanism to inject the identity condition into the intermediate representations of the diffusion model. The introduction of partial dropout in the components of the identity context during training to prevent overfitting and increase intra-class diversity was another advancement. The authors achieved state-of-the-art performances, in mainstream datasets e.g. LFW. The pipeline is demonstrated in Figure 3.7.

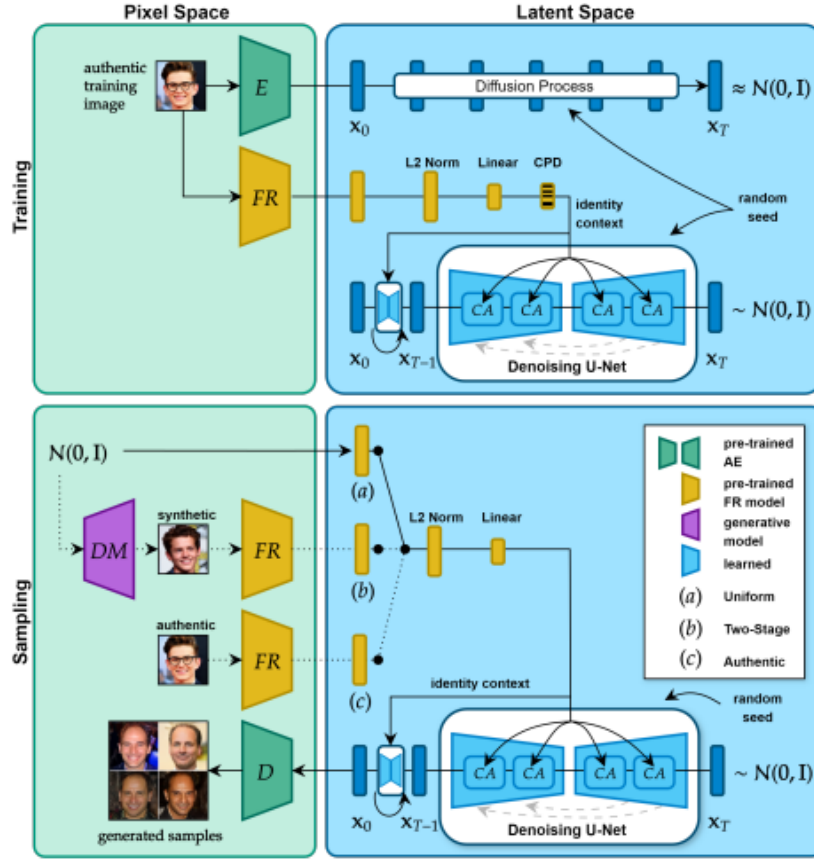


Figure 3.7: The illustration is divided into two segments. The upper segment details the training process, where a denoising U-Net is guided by contextual information derived from a pre-trained Face Recognition model’s features. This training occurs within the latent domain of a pre-trained Autoencoder, with the diffusion process providing targets for the reverse learning sequence. The lower segment explains the sample creation process, where the trained Diffusion Model can generate samples using three identity contexts: real, two-stage, or synthetic uniform. By maintaining a fixed identity context and modifying the noise, diverse samples for the same identity are achievable. From Boutros et al. (2023a).

Geissbühler et al. (2024) proposed a physics inspired approach, to generate the intraclass and interclass variations. To sample a distribution of synthetic samples, they proposed a method called inspired by the Langevin equation, that improves in an interactive way the set of latent vectors, aiming for an optimally distributed synthetic latent vectors. Firstly, they extract the synthetic samples embeddings, that were generated by random sampled latent vectors. They introduce two quadratic loss functions: the first, inspired by granular mechanics, repels embeddings up to a certain threshold, while the second pulls latent vectors towards the generator’s average latent vector. This process iteratively increases inter-class embedding distances while maintaining a compact latent space distribution for high-quality image generation. For the intraclass variation, they proposed the Dispersion algorithm, that works similar to Langevin, but works on the latent space. Another quadratic loss function is added to maintain the embeddings of the variations close to a certain threshold. They also enhance the initialization procedure adding a random linear combination of Covariate vectors before the first iteration. The solution is described at Figure 3.8.

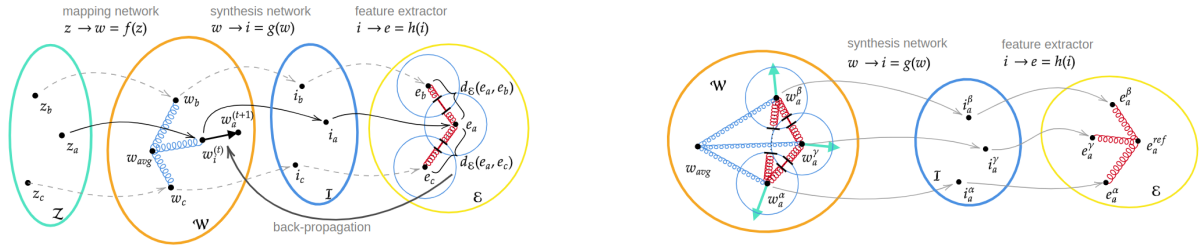


Figure 3.8: For each identity a , a random vector z_a is sampled and mapped to initial latent $w_a^{(0)}$. An image i_a is generated from w_a and its embedding e_a computed. Loss functions on embedding space E and latent space W are used to update latents $w_a^{(t)} \rightarrow w_a^{(t+1)}$ over N_{iter} iterations. This method is used to generate the interclass variations (left figure). For each identity a and variation α , initialize latent $w_a^{\alpha(0)}$ near w_a^{ref} (right figure). Losses pull embeddings towards e_a^{ref} , latents towards average w_{avg} , and repel nearby latents. Repeat for all a and α over N_{disp_iter} iterations. This method is used to generate the intraclass variations. From Geissbühler et al. (2024).

HyperFace, proposed by Shahreza and Marcel (2024), is a novel approach to generating synthetic face recognition datasets by framing the dataset creation as a packing problem within the embedding space of a face recognition model, represented on a hypersphere. This method formalizes the packing problem as an optimization task, solved using a gradient descent-based approach, and employs a conditional face generator to synthesize face images from the optimized embeddings. The resulting synthetic datasets are used to train face recognition models. The explanation are present at the Figure 3.9.

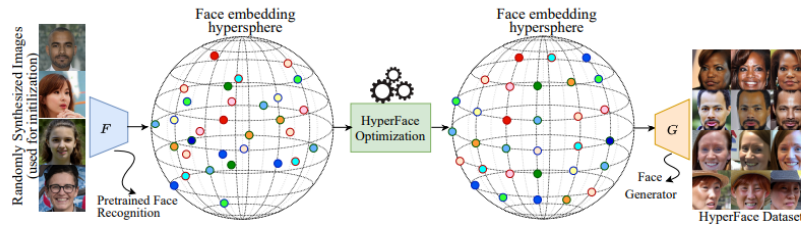


Figure 2: Block diagram of HyperFace Dataset Generation

Figure 3.9: The images are firstly generated with a synthesizer, its embeddings are computed, and the optimization process is applied. As a result, the interclass variation is achieved and the embeddings are used to generate the final dataset images. From Shahreza and Marcel (2024).

In contrast, Vec2face, elaborated by Wu et al. (2024), is composed of a feature-masked encoder decoder architecture. Using vectors with low similarity as input, different identities can be generated. Also, by weekly perturbing the identity vector, the intraclass variations are applied. The method illustration is presented in Figure 3.10. They also proposed a gradient descent method, that adjusts the vector values to generate images with designated attributes.

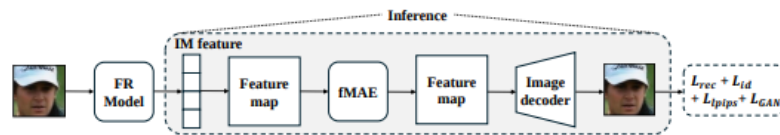


Figure 3.10: The IM feature is computed by a facial recognition model, and then it is expanded into a feature map. The latter is processed by a feature mask autoencoder (fMAE), where the rows are randomly masked. Then they are fed to an image decoder, that reconstructs the pixels. The training is conducted calculating the MSE and cosine similarity between the reconstructed image and the ground truth, and the perceptual loss and GAN loss are used to ensure a correct facial structure and increase the sharpness of the generated images. From Wu et al. (2024).

Another approach, called Arc2face (Papantoniou et al., 2024), builds upon a stable diffusion model, and adapted it to the ID generation, conditioned to the identity embeddings. They focuses exclusively on ID vectors derived from ArcFace, a prominent face recognition model. This approach allows the model to generate images that maintain strong identity consistency without the need for textual input. The method is described in Figure 3.11.

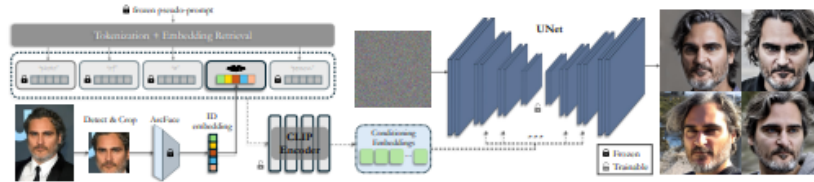


Figure 3.11: The Arcface embedding are concatenated to the CLIP input, that outputs conditional embeddings that are used for cross-attention control. The UNet and the encoder are optimized using a million scale dataset, and further finetuned on a high resolution dataset, without any text annotations. From Papantoniou et al. (2024).

VairFace (Yeung et al., 2024) is another method, that also builds upon diffusion models and CLIP models to generate the images and define the demographic labels, respectively. They also refine the demographic labels using a face recognition model, applied afterward. For generating the interclass diversity, Face Vendi Score Guidance is integrated to the diffusion loss. To balance the intraclass identity preservation and diversity trade-off, they proposed the Divergence Score Conditioning. A more detailed explanation about the method pipeline is presented in Figure 3.12.

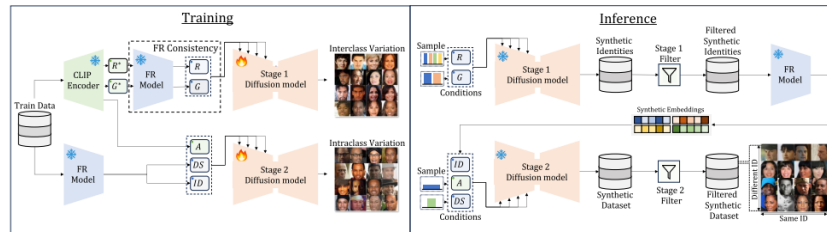


Figure 3.12: The training predictions for race (R^*), gender (G^*) and age (A) are outputs of a CLIP model. Next to ensure Facial recognition consistency, its used a FR pipeline to refine the race and gender labels, as well as computing the ID and divergence (DS) labels. These labels are used to train the diffusion models in stage 1 and 2. The inference procedure consists of using the balanced synthetic identities provided in stage 1, filtered and processed, to generated synthetic embeddings. These and randomly sampled A and DS are used as input to the second stage diffusion model to generate the synthetic dataset, that a passed to the second filtering stage to create the filtered dataset. From Yeung et al. (2024).

Kim et al. (2024) developed VigFace, that proposes pre-assigning virtual identities in the feature space. The authors trained using real and virtual prototypes using a prominent loss function (Arcface) and generated a feature space for both real and virtual identities. Subsequently, they generated the synthetic identities using a diffusion model, by synthesizing virtual entities from devised virtual prototypes. The method description is illustrated in Figure 3.13.

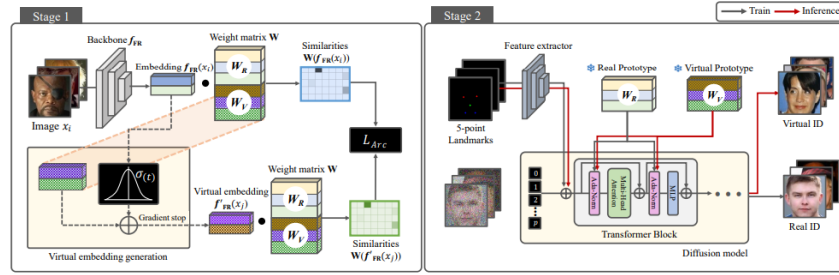


Figure 3.13: The pipeline was trained with real prototypes indicated as WR , and k prototypes for virtual IDs, denoted as WV . The virtual embedding $f'_{FR}(x_j)$ is designed to simulate the distribution of real embeddings. Subsequently, a diffusion model is used to generate the synthetic images based on the virtual prototypes. From Kim et al. (2024).

In respect to the same topic, new proposals were submitted to a competition event called the Face Recognition Challenge in the Era of Synthetic Data (Melzi et al., 2024), where the participants faced the challenges of training a pipeline with synthetic data using a combination of DCFace (Kim et al., 2023) and GANDiffFace (Melzi et al., 2023) datasets and also using a combination of real datasets and the synthetic ones on other sub task. It was asked to reduce the synthetic to real performance gap and also contribute to the reduce the discrepancy in racial and gender bias in facial recognition technology. As summary, the winners solutions proposes pipelines trained with Adface (Kim et al., 2022) and Arcface (Deng et al., 2019a) loss functions using a resnet (He et al., 2016) as backbone. There are also similar competition events, called Synthetic Data for Face Recognition (SDFR) and a extension of the FRCSyn challenge.

The Table 3.1 summarizes all the previously described methods, and adds more detailed information, such as the generator training dataset names and the number of images used to train, generator type and name given, if it was trained with id labels (important due to competitions that do not allowed id labels during training the generator), the number of images generated (based on the selected dataset between the provided ones) and a brief method description.

Dataset	Images (IDs x imgs/ID)	Generator Dataset	Training	Generator	Name of the Generator	Trained with ID Labels	Number of Images Used to Train the Generator	Method
Dcface	1.3M (40K x 5 + 20K x 55)	FFHQ (step 1) and CASIA-Webface (step 2)		Diffusion Model	Not named	Yes	70000 images (first stage) + 490414 images (second stage), 2-stage diffusion module consisting of a sampling stage and a mixer stage.	2-stage diffusion model consisting of a sampling stage and a mixer stage.
Synface	1M (10K x 100)	Flickr-Faces HQ (FFHQ)		GAN	DiscoFaceGAN	No	70000 imgs	Applied identity and domain mixup techniques, sampling from a style conditioning model.
Digiface	499995 (99999 x 5)	Not named		3D model	Not named	No	511 face scans	3D model applied to generate images posteriorly heavily augmented.
GANDiffFace	501172 (9305 x aprox. 53.86)	FFHQ (StyleGanV3 training), StyleGAN generated images and DreamBoth generated images, both for fine tuning		GAN + Diffusion model	StyleGAN3 + DreamBoth	Yes	70000 images (StyleGAN training) + finetune DreamBoth for each identity using 6 same class images and 200 persons images for regularization	The StyleGAN images were filtered to meet the target demographic distributions, and serve as input to a diffusion model to generate intraclass variations.
SFace	1885877 (10572 x aprox. 178.38)	Flickr-Faces HQ (FFHQ)		GAN	StyleGAN2-ADA	Yes	70000 imgs	Dataset generated using random latents imputed into a StyleGAN-ADA
Idnet	1057200 (10572 x 100)	Flickr-Faces HQ (FFHQ)		GAN	StyleGAN2-ADA	Yes	70000 imgs	Enhanced a standard GAN framework, by adding a third player designed to ensure that identities are separable.
IDiff-Face	502500 (10050 x 50)	Flickr-Faces HQ (FFHQ)		Diffusion model	vq-f4 autoencoder	No	400M CLIP-filtered image-text pairs (pre-training) + 70000 imgs	A diffusion model is conditioned on identity context to produce identity separable images.
DisCo	1920000 (30K x 64)	Flickr-Faces HQ (FFHQ)		GAN	StyleGAN2	No	70000 imgs	Uses brownian motion of particles to generate interclass and intraclass variations.
Hyperface	640K (10K x 64)	LAION-400M(pre-training) + Webface42m(training), FFHQ and CelebA-HQ (both for fine tuning)		Diffusion model	Stable-diffusion-v1-5	No	400M CLIP-filtered image-text pairs (pre-training) and 42M imgs for training and 1000000 imgs for fine tuning	An optimization algorithm is used to generate embeddings of identities well spread across a hypersphere, further used to serve as guidance for the Arc2face method.
Vec2Face	1M (20K x 50)	Webface4M subset		Decoder + GAN based training	Not named	No	1M images for training	Feature-masked encoder decoder that uses a sampled vector as input and controls the face images and their attributes.
Arc2Face	1.2M (20K x 50 + 40K x 5)	LAION-400M(pre-training) + Webface42m(training), FFHQ and CelebA-HQ (both for fine tuning)		Diffusion model	Stable-diffusion-v1-5	No	400M CLIP-filtered image-text pairs (pre-training) and 42M imgs for training and 1000000 imgs for fine tuning	Diffusion model guided by identity features generated by a CLIP text encoder.
Vairface	1.2M (60K x 20)	CASIA-Webface		Diffusion model	Hourglass Diffusion Transformer (HDT)	No	490414 images	2-stage diffusion model, guided by identity features (i.g. gender and age) produced by ViT-L-14 MetaCLIP model and refined by IResnet-100 model.
Vigface	1.2M (60K x 20)	CASIA-Webface		Diffusion model	DiT-B	No	490414 images	Pre-assigning virtual identities in the feature space and guiding the DiT-B using the virtual identities.

Table 3.1: Table summarizing the objectives and details of facial recognition datasets.

In the exploration of recent studies in facial recognition, it is evident that the field has made significant strides, particularly with the integration of large-scale datasets and advanced learning algorithms. Extensive databases like MS1MV2, Glint360k, and Webface260M have been pivotal in enhancing face verification accuracy, enabling the development of more refined and capable models. The innovation in loss functions such as ArcFace, CosFace, and AdaFace, along with the adoption of deeper network architectures, has significantly improved feature extraction and classification capabilities. Despite these advancements, challenges remain, particularly concerning data variations in pose, aging, and expressions, and ethical concerns related to privacy and demographic representation. These challenges highlight the need for continued focus on improving training methodologies and data generation techniques, especially through synthetic data to address current limitations.

Moving into the methodology section, it is outlined a series of reproducible strategies for image preprocessing, training configurations, and evaluation protocols designed to maximize the performance of facial recognition systems, reflecting the best practices derived from the reviewed literature.

4 METHODOLOGY

This chapter describes the methodology used to perform the proposed experiments. The section 4.1 describes the image preprocessing. The section 4.2 describes the training choices, including the backbone used, loss function applied, augmentation used, training parameters and the training datasets, resulted from the studied methods from section 3. The sections 4.3 describe the evaluation datasets and the metrics used for evaluation.

4.1 IMAGE PREPROCESSING

Before training, it is necessary to crop and align images using a landmark detector. For this work, it was employed the RetinaFace facial landmark detector, developed by Deng et al. (2019b). Using the five landmark points detected, the face is cropped to a center region, and it is performed alignment to all training images.

Additionally, normalization is applied, which is the process of scaling the pixel values of an image to a specific range, typically to improve convergence during training and the performance of the model. Normalization involves adjusting pixel values to have a mean of zero and a standard deviation of one.

4.2 TRAINING CHOICES

4.2.1 Used Backbone

One of the most impressive forms of Artificial Neural Network architecture is the CNN. According to Arandjelovic et al. (2016), in recent years, such networks have emerged as protagonists in category recognition tasks such as object classification, scene recognition or object detection. The basic principles of CNNs were introduced during the 1980s by LeCun et al. (1989, 1998) and their recent success is largely due to advances in the computational power of GPUs in conjunction with the availability of large labeled databases.

A very popular used architecture for facial recognition are the ResNets (He et al., 2016). These networks are able to learn low/medium/high level features when increasing in depth, respectively. This is possible due to the use of skip connections, which alleviate the vanishing gradient problem encountered in training deep neural networks.

The chosen backbone for performing the task was the iResNet-101, introduced by Duta et al. (2021), and one of the top-performing backbones for deep FR, according to Deng et al. (2021). The authors improved the vanilla ResNet by changing the arrangement of the components and subdividing the building blocks into three stages with the aim of improving the flow of information through the network. They also introduced a projection shortcut that reduces information loss and a convolution block that operates in a larger number of channels, improving performance as this block is the only component responsible for learning spatial patterns. Applying these changes provided consistent improvements in accuracy and training convergence over the baseline.

4.2.2 Loss Function

The applied loss function is the ArcFace loss, developed by Deng et al. (2019a). A common loss function used for facial recognition is the softmax loss. This function is also called cross-entropy

loss, a commonly used loss function in classification tasks in machine learning. It is often used in conjunction with softmax activation in the output layer of a neural network. The softmax function takes a vector of real-valued scores (logits) and normalizes it into a probability distribution over multiple classes. For classification, after applying softmax to the output layer of a neural network, the resulting probabilities represent the model's confidence in each class. Despite being useful for closed-set classification problems, it is not discriminative enough for open-set facial recognition problems. This loss is presented in equation 4.1.

$$L = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{W_{y_i}^T \cdot x_i + b_{y_i}}}{\sum_{j=1}^N e^{W_j^T \cdot x_i + b_j}} \right) \quad (4.1)$$

The deep feature of the i -th example is represented by the variable $x \in \mathbb{R}^{512}$. The $W_j \in \mathbb{R}^{512}$ represents the j -th column of $W \in \mathbb{R}^{d \times n}$. The bias term is represented by $b_j \in \mathbb{R}^n$, while N and n represent the batch size and class number.

The Arcface loss (Deng et al., 2019a) is defined by equation 4.3. This is a result of rewriting the softmax in equation 4.1 by using the fact that $W_j^T x_i = \|W_j\| \cdot \|x_i\| \cdot \cos(\theta_j)$. By fixing $b_j = 0$ and using $L2$ normalization to $\|W_j\| = 1$ and $\|x_i\| = s$, where s represents a hyper-sphere with radius s , see equation 4.2. After this, m is the added angular margin to the feature class softmax output. This margin results in improved intra-class consistency and inter-class separation.

$$L = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{s(\cos(\theta_{y_i}))}}{e^{s(\cos(\theta_{y_i}))} + \sum_{j=1, j \neq y_i}^C e^{s \cos \theta_{ji}}} \right) \quad (4.2)$$

$$L = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^C e^{s \cos \theta_{ji}}} \right) \quad (4.3)$$

4.2.3 Augmentation Applied

To the training data, it was applied three different augmentations: Random Horizontal Flip, RandAugment, and Random Erasing. The Random Horizontal flip technique refers to the process of randomly flipping an image along its horizontal axis. By randomly flipping images, the model is exposed to different perspectives, effectively increasing the size and diversity of the training dataset without the need for additional data collection. Models trained with augmented data are less likely to overfit and more likely to generalize well to unseen data, helping the model become invariant to certain transformations and making it more robust to variations in the input data.

RandAugment, introduced by Cubuk et al. (2020), is a data augmentation technique introduced to simplify and improve the process of augmenting images for training deep learning models. This method aims to enhance the performance of models by applying a set of random transformations to the training images, thereby increasing the diversity of the training data and improving the model's robustness and generalization capabilities. Unlike other augmentation techniques that require a search over a large space of possible augmentations, RandAugment simplifies the process by reducing the number of hyperparameters to tune. There are two hyperparameters: the number of augmentation transformations to apply sequentially to each image (N), and the magnitude or intensity of the transformations (M). This technique uses a fixed set of predefined transformations such as identity, autoContrast, equalize, rotate, solarize, color, posterize, contrast, brightness, sharpness, shear-x, shear-y, translate-x, translate-y.

Random erasing involves randomly selecting a rectangular region within an image and erasing its pixel values, effectively introducing occlusions or noise. The pixel values within the selected region are either set to a constant value (e.g., zero) or replaced with random values. The size and aspect ratio of the erased region are controlled by specific parameters, allowing for flexibility in the augmentation process. By introducing occlusions, the model learns to focus on the most discriminative parts of the image, making it more robust to partial occlusions and noise in real-world scenarios.

For both augmentations, it uses the PyTorch transforms default parameters, applied sequentially. The random horizontal flip can also be applied. In general, all the augmentations combined enhance the model robustness and generalizability, resulting in accuracy gains.

4.2.4 Method Description

It was employed ArcFace as the loss function and iResNet101 as the backbone. These were chosen, because is one of the top-performing models for deep FR (Deng et al., 2021). As dicussed in section 4.2.3 images used for training were augmented. The models were trained using the Insightface library. It was used the SGD optimizer, setting momentum to 0.9 and weight decay to 5×10^{-4} . The learning rate was set to 0.02 and decayed at each iteration following the equation 4.4.

$$\left(\frac{1.0 - \frac{l}{t}}{1.0 - \frac{l-1}{t}} \right) \quad (4.4)$$

The variables l and t represent the current iteration and the total number of iterations, respectively. The model was trained for 20 epochs within a batch size of 128, on an NVIDIA TITAN Xp with 12GB memory.

It was applied the same training configurations for all the studied works, despite VIGFace, VariFace and Hyperface, that the results were taken from the article itself, as their dataset are not yet publicly available. As a disclaimer, it was selected and trained datasets using a criteria, which is the best accuracy for Iddif-Face, DCface, Digiface and Disco. In the case of IDiff-Face, the selected database was the Uniform version, and for DCFace, the version with 1.3M images. For Digiface and Disco, the one with the largest number of identities was selected. Also, for the datasets Arc2face and Vec2face, the dataset with the number of images that are close to the other datasets were taken. For the remaining ones, only one option was provided and used to train, or the results were taken from the article itself, and adopted by the study.

4.3 EVALUATION PROTOCOL

For each selected dataset, it was performed a 10-fold cross verification test using the selected datasets (i.e. LFW, CPLFW, CFPFP, CALFW, AGEDB, IJBB, IJBC). For the TinyFaceR1 and TinyFaceR5, an identification protocol is performed, that is a one-to-many matching problem.

The Table 4.1 describes the datasets used for evaluation, the evaluation protocol and their chalanges.

The metric used to evaluate the mainstream datasets (i.e. LFW, AGE, CFPFPFP, CPLFW, CALFW) is the best accuracy. For the IJBB and IJBC datasets, the metric is the TPR at an FPR of 0.01%. For the TinyFaceR1 and TinyFaceR5, the rank-1 and rank-5 accuracy is measured. It was also used the ROC curves to evaluate the mainstream datasets, and IJBB and IJCB. These curves were calculated using the function `roc_curve` from `scikit-learn`.

Database	Evaluation Protocol	Challenges
LFW	3000 genuines and 3000 impostors	Uncontrolled conditions, varied poses and expressions
AGEDB	3000 genuines and 3000 impostors	Aging effects, expressions, pose variations
CFPFP	3500 genuines and 3500 impostors	Frontal-profile mismatches, pose variations
CPLFW	3000 genuines and 3000 impostors	Distinct poses, cross-pose matching
CALFW	300 genuines and 3000 impostors	Cross-age facial recognition, aging effects, varied expressions, and poses
IJBB	10,270 genuines and 8,000,000 impostors	Variations in pose, illumination, image quality, low false positive rates
IJBC	19,557 genuines and 15,638,932 impostors	Large-scale data, uncontrolled conditions, pose variance, low false positive rates
TinyFaceR1	5,139 labelled facial identities	Low resolution FR at large scales, variations in occlusion and pose
TinyFaceR5	5,139 labelled facial identities	Low resolution FR at large scales, variations in occlusion and pose

Table 4.1: Summary of facial Recognition datasets, evaluation protocols, and their challenges.

In this chapter it was described the training methodology used to train the publicly available dataset and described in the studied works (section 3). The evaluation is also described, by evidencing the metrics and curves used to evaluate each dataset, on test datasets.

5 RESULTS AND DISCUSSION

This chapter describes the results and discussion achieved in the comparative study. A discussion about their advantages and disadvantages are provided.

The achieved results, for each test dataset and corresponding used metric to evaluate, are presented in Table 5.1 and Table 5.2. The ROC curves for all verification sets are presented in Figures 5.1, 5.2 and 5.3. For VIGFace, VariFace, and Hyperface, the results were taken from the article itself, as their dataset are not yet publicly available. This makes a difference, since the training methodology is different, but for the sake of completeness, the original training results were included. Also, as the articles do not report the results in the IJBB, IJBC, TinyFaceR1, and TinyFaceR5 the mentioned methods were omitted from Table 5.2.

Dataset	LFW	CPFLW	CFPPF	CALFW	AGEDB	AVG
Webface4m (real)	99.81	94.68	98.50	95.91	97.48	97.28
Arc2Face	99.48	92.53	97.57	95.21	95.18	95.99
VairFace(*)	99.45	90.63	95.61	94.13	94.75	94.91
CASIA-WebFace (real)	99.25	89.65	97.07	93.33	94.40	94.74
VIGface(*)	99.15	88.88	96.66	92.22	92.73	93.93
DCFace	98.93	87.03	93.00	92.93	92.55	92.89
Vec2Face	98.83	87.20	91.08	93.18	93.30	92.72
HyperFace(*)	98.73	85.43	89.54	90.05	87.52	90.25
DisCo	99.03	76.53	84.17	92.98	91.60	88.86
IDiff-Face	97.31	74.50	79.12	85.63	77.78	82.87
Digiface	94.38	74.38	80.97	76.06	71.20	79.40
GANDiffFace	94.06	74.38	78.44	78.30	68.28	78.69
Idnet	92.58	73.48	76.08	77.13	67.96	77.45
Sface	92.52	72.33	73.57	76.66	70.28	77.07
Synface	81.36	61.78	65.85	64.10	60.66	66.75

Table 5.1: Performance result across selected datasets on mainstream datasets. * indicates that the results were taken from the original article itself. The mainstream dataset are called Labelled Faces in the Wild (LFW), Cross-Pose LFW (CPFLW), Celebrities in Frontal-Profile (CFP) (protocol FP), Cross-AgeLFW (CALFW), Age Database (AGEDB).

An upper bound for the synthetic datasets compared is Webface4M (Zhu et al., 2021), which is a subset version of Webface42M that has approximately 42 million images of 2M identities, while Webface4M contains 4M images from 200K identities. It is an upper bound because a real dataset and a better performance is obtained when training with this dataset in all dataset protocols. It is also possible to verify this behavior on the ROC curves (Figures 5.1, 5.2, 5.3), for the verification datasets, where the yellow line (representing the TPR at fixed FPR operational points), is always above the other lines in the graph, despite one occurrence where the Arc2face method performed better on points very close to zero FPR on IJBB (Figure 5.3), but further Webface4M surpassed it.

The Arc2face dataset comes in sequence, achieving an avg accuracy of 95.99 on mainstream datasets and 79.47 on the IJB two variants and TinyFaceR1,R5, Tables 5.1, 5.2 respectively. The combination of diffusion backbone and effectively transform the text encoder into a face encoder specifically tailored for projecting ArcFace embeddings into the CLIP latent

Dataset	IJBB@0.01	IJBC@0.01	TinyFace R1	TinyFace R5	AVG
Webface4m (real)	95.18	96.67	73.81	76.52	85.55
Arc2Face	89.10	92.66	65.71	70.41	79.47
CASIA-WebFace (real)	82.55	86.66	57.59	63.65	72.61
DCFace	79.79	83.56	53.88	60.13	69.34
Vec2Face	62.22	56.16	57.48	63.49	59.84
DisCo	59.24	62.01	51.15	58.20	57.65
IDiff-Face	48.90	50.27	45.86	54.85	49.97
Digiface	35.92	41.17	55.31	62.98	48.85
Idnet	45.46	49.69	44.44	54.80	48.60
Sface	8.92	4.59	35.30	43.88	23.17
Synface	0.19	0.18	45.54	54.58	25.12
GANDiffFace	0.61	0.54	39.80	45.81	21.69

Table 5.2: Performance result across selected datasets on IARPA Janus Benchmark-C (IJBC), IARPA Janus Benchmark-B (IJBB), TinyFace R1, and TinyFace R5.

space, making this a strong solution to the problem. It is also scalable, capable of generating a large number of images (i.e. the released dataset contains 21M facial images from 1M identities at a resolution of 448×448). However, it has its downsides, like the use of large amount of real data (i.e. 42M imgs for training and 1M imgs for fine tuning) the text encoder (Radford et al., 2021) and the diffusion backbone (Rombach et al., 2022).

Due to that, the method falls into a category where it was not properly made for dealing with the concerns regarding the use of web scrapped datasets, but for facial attribute augmentation. This method was considered in the evaluation because it is an allowed method for submission to competitions events (Melzi et al., 2024; DeAndres-Tame et al., 2024; Shahreza et al., 2024).

VairFace is 2-stage diffusion model, guided by identity features (e.g. gender and age) produced by ViT-L-14 MetaCLIP model (Xu et al., 2023) and refined by IResnet-100 (Duta et al., 2021). They achieved an avg accuracy of 94.91 on the mainstream datasets (Figure 5.1). They also sample identities considering a more equitable demographic dataset. This is very relevant, since it is already known that celebrity datasets also have imbalanced racial distribution (e.g., 84.5% of the faces in CASIA-WebFace are Caucasian faces), leading to non equitable recognition accuracy for the under-represented racial groups. They improved over the real dataset that it was originally trained on, that is the CASIA-WebFace, just losing in the CFPFP protocol 95.61 vs. 97.07, probably due to limited number of profile images generated. Considering the Webface4m as an upper bound, the synthetic to real gap is 2.37.

Vigface proposed pre-assigning virtual identities in the feature space and guiding the DiT-B (Peebles and Xie, 2023) using the virtual identities. The performance is closed to the real dataset that the method was trained (i.e. CASIA-WebFace), with a gap of 0.81, probably lower if trained with the methodology proposed in this work and described in section 4.2 (i.e. ResNet-100 with Arcface loss function and augmented data).

The next evaluated dataset is DCface, a 2-stage diffusion model consisting of a sampling stage and a mixer stage. It achieves an average accuracy of 92.89 on mainstream datasets and 69.34 on the IJBB, IJBC and TinyFaceR1,R5 (Tables 5.1 and 5.2). Compared to the dataset that the generator was trained on (i.e. CASIA-WebFace) the gap on mainstream dataset is of 1.85, while the gap IJB and TinyFace families is of 3.27. Considering the ROC curves (Figures 5.1, 5.2 and 5.3), the method curve followed very closely the CASIA-Webface one, just in the CFPFP protocol the difference were bigger. They also considered sampling a balanced demographically

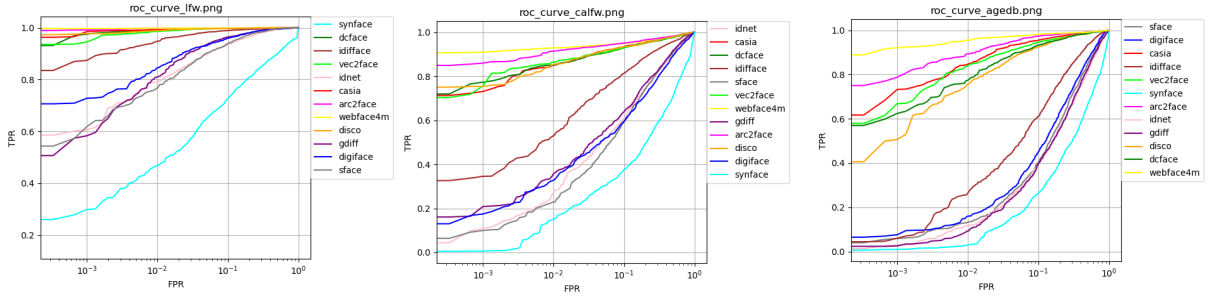


Figure 5.1: Achieved ROC curves on LFW, CALFW and AGEDB for the trained datasets. TPR means True Positive Rate and FPR means False Positive Rate

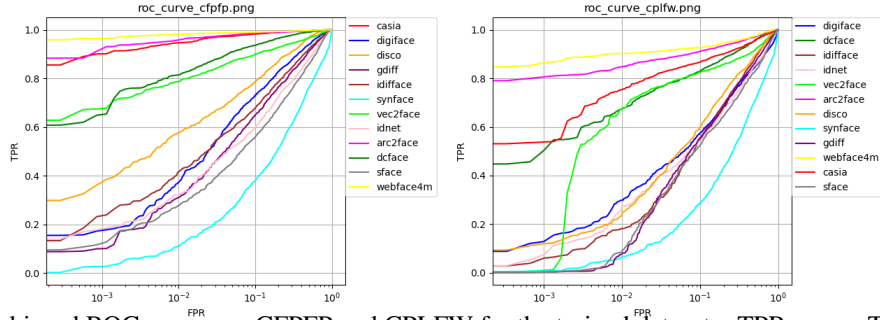


Figure 5.2: Achieved ROC curves on CFPFP and CPLFW for the trained datasets. TPR means True Positive Rate and FPR means False Positive Rate

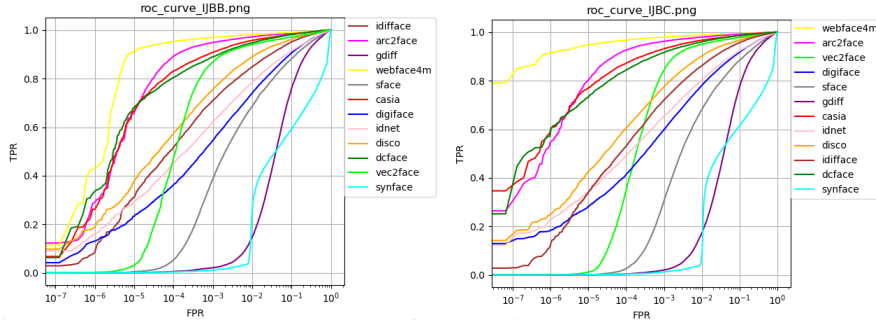


Figure 5.3: Achieved ROC curves on IJBB and IJBC for the trained datasets. TPR means True Positive Rate and FPR means False Positive Rate

distributed dataset. This method was allowed at the FRCSyn 1st and 2nd editions, but not allowed in the SDFR competition (Shahreza et al., 2024), because it was trained using id labels.

Vec2face proposed a unique solution to the problem, that is a feature-masked encoder decoder that uses a sampled vector as input and controls the face images and their attributes. They trained their method on a subset of Webface4M, with 1M images, achieving an average accuracy of 92.72 and 59.84 on the mainstream and IJBB, IJBC and TinyFaceR1, TinyFaceR5 (Tables 5.1 and 5.2). It is also possible to verify that the ROC curves (Figures 5.1, 5.2 and 5.3) is not well behaved in comparison to DCFace and CASIA-Webface.

In sequence comes Hyperface, an optimization algorithm that is used to generate embeddings of identities well spread across a hypersphere, further used to serve as guidance for the Arc2face method. However, it is possible to verify that those embeddings are less effective than the PCA generated ones (i.e. embedding sampling method for Arc2face method). This has an upside that the learned embeddings are less related in similarity to the real data as the PCA generated ones, because PCA are inherently linked to the actual data provided for the

analysis. They achieved an average accuracy of 90.25 on mainstream datasets, and a gap to CASIA-Webface of 4.49 (Table 5.1).

DisCo comes in sequence, and is based on the brownian motion of particles to generate interclass and intraclass variations. This method achieved an average accuracy of 88.86 on mainstream datasets and an average of 57.65 on IJBB, IJBC and TinyFace, configuring a gap to CASIA-Webface of 5.88 and 14.96 respectively (Tables 5.1 and 5.2). Due to the use of a GAN, named StyleGAN2 and a low training data regime (ie. FFHQ), they achieved worst results when compared to diffusion models, due to characteristics inherent to these models.

IDiff-Face comes in sequence, which is a diffusion model conditioned on identity context to produce identity separable images. They achieve an average accuracy of 82.87 on mainstream datasets and an average of 49.97 on the IJBB, IJBC and TinyFace protocols, bridging the synthetic-to-real accuracy gap to 12.1 and 22.64 considering CASIA-Webface on the respectively group of datasets (Table 5.1 and 5.2). This result reveals the potential that diffusion models can have when generating images with variations in pose, age, expression, and illumination, containing unique information. As stated in Kim et al. (2023), the diffusion models can generate a bigger number of unique identities than GAN's when the sample approach is random or guided by identity features, which results in a dataset with more variety. Also, they use a pre-trained model to extract embeddings that are used by the conditional generator model based on diffusion. However, as identity labels were not used to train the face generator model, this dataset was allowed in the SDFR competition and was adopted by the majority of the competitors.

Digiface proposed a unique solution to the problem, by using a computer graphics pipeline to render the facial images. With this, they were able to achieve an average accuracy of 79.40 on mainstream datasets and an average of 48.85 on the IJBB, IJBC and TinyFace protocols, bridging the synthetic-to-real accuracy gap to 15.34 and 23.76 considering CASIA-Webface on the respectively group of datasets (Table 5.1 and 5.2). They achieve a higher value on the CFP-FP dataset when compared to IDiff-Face, which indicates that profile images are more present in the training dataset of this solution. However, this approach is extremely computationally costly and might not be available for research. This solution also has the credit of not using large-scale real face datasets to train some components of their pipeline, leaving unresolved ethical problems. However, at its core, the challenge of creating complex synthetic data that closely mirrors certain authentic references is akin to the chicken-and-egg dilemma. It may be beneficial to reconceptualize this issue to highlight that synthetic data inherently relies on genuine prior knowledge. Instead, emphasis should be placed on devising methods that ensure the original data remains difficult to reconstruct, while also ensuring that the synthetic dataset effectively captures a good enough representation of the reality (Geissbühler et al., 2024).

GANDiffFace face dataset resulted in an average accuracy of 78.69 on mainstream datasets and an average of 21.69 on the IJBB, IJBC and TinyFace protocols (Tables 5.1 and 5.2). They used StyleGAN3 to generate the images and further input those images into a diffusion model (i.e., Dream Booth), to generate the intra-class variations. The additional diffusion model enables the dataset to have a more realistic intra-class variation. The authors also considered demographic equitable sampling. However, this comes at cost of finetuning the diffusion model for each identity generated. Also, the dataset do not main its position on the IJBB,IJBC and TinyFace protocols, having difficult on operating on fixed FMR points and identifications scenarios.They applied demographic equitable sampling. SFace, that is a dataset generated using a class conditioned StyleGAN2-ADA achieved an average accuracy of 77.07 on mainstream datasets and an average of 23.17 on IJBC,IJBB and TinyFace protocols (Tables 5.1 and 5.2).

An upgrade of Sface is IdNet, which achieved an average accuracy of 77.07 on mainstream datasets and an average of 23.17 on IJBC,IJBB and TinyFace protocols (Tables 5.1

and 5.2). According to the authors, “SFace suffers from relatively low identity separability which might lead to less optimal face verification accuracies when such synthetic data is used to train FR” (Kolf et al., 2023). To deal with this they integrate to the GAN min-max game and identity separable loss, named ID3, and a domain adaptation loss to make the generator learn to identify information encoded and generate more identity separable images. However, this comes with the cost of using identity labels in training the generative framework.

Lastly, comes Synface, which employed DiscoFaceGAN (Deng et al., 2020) and identity mixup and domain mixup techniques. The solution achieved an average accuracy of 66.75 on mainstream datasets and an average of 25.12 on IJBC, IJBB and TinyFace protocols (Tables 5.1 and 5.2). The lower performance can be a result of the generator, that provides few unique samples. Also, this work has the credit of being, one of the first to generate a synthetic dataset to train a facial recognition model.

6 CONCLUSION

In summary, synthetic data generation for facial recognition has seen notable advancements with methods like diffusion models, Generative adversarial networks, and 3D rendering techniques that aim to mimic the diversity and complexity of real-world datasets. Techniques such as DCFace’s diffusion model and Synface’s mixup methods have started closing the performance gap between synthetic and real data, improving accuracy while addressing ethical concerns associated with using real identities. These approaches focus on enhancing intra-class variations and leveraging synthetic data diversity to build robust facial recognition models without relying on real-world data.

However, significant challenges remain, particularly regarding overfitting, computational costs, and achieving proper demographic representation in synthesized datasets. Models like HyperFace and Vec2Face have experimented with different strategies to ensure diversity and authenticity, yet replicating the high performance of models trained with real data remains an issue. Despite innovations like VairFace and Arc2Face’s identity-focused diffusion applications, synthetic datasets still face limitations in realism and representation. Future research must focus on reducing computational demands, maintaining ethical standards, and ensuring that synthetic datasets effectively capture a wide range of demographic features to become an alternative to real data in training facial recognition systems.

REFERENCES

- Ahonen, T., Hadid, A., and Pietikäinen, M. (2004). Face recognition with local binary patterns. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I* 8, pages 469–481. Springer.
- An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., et al. (2021). Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449.
- Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307.
- Bae, G., de La Gorce, M., Baltrušaitis, T., Hewitt, C., Chen, D., Valentin, J., Cipolla, R., and Shen, J. (2023). Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535.
- Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720.
- Boutros, F., Grebe, J. H., Kuijper, A., and Damer, N. (2023a). Idiff-face: Synthetic-based face recognition through fuzzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661.
- Boutros, F., Huber, M., Siebke, P., Rieber, T., and Damer, N. (2022). Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE.
- Boutros, F., Struc, V., Fierrez, J., and Damer, N. (2023b). Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, page 104688.
- Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.
- DeAndres-Tame, I., Tolosana, R., Melzi, P., Vera-Rodriguez, R., Kim, M., Rathgeb, C., Liu, X., Morales, A., Fierrez, J., Ortega-Garcia, J., et al. (2024). Frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3183.
- Deng, J., Guo, J., An, X., Zhu, Z., and Zafeiriou, S. (2021). Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444.
- Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019a). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

- Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. (2019b). Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*.
- Deng, Y., Yang, J., Wen, D. C. F., and Tong, X. (2020). Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Déniz, O., Castrillon, M., and Hernández, M. (2003). Face recognition using independent component analysis and support vector machines. *Pattern recognition letters*, 24(13):2153–2157.
- Duta, I. C., Liu, L., Zhu, F., and Shao, L. (2021). Improved residual networks for image and video recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE.
- Geissbühler, D., Shahreza, H. O., and Marcel, S. (2024). Synthetic face datasets generation via latent space exploration from brownian identity diffusion. *arXiv preprint arXiv:2405.00228*.
- Grother, P., Grother, P., Ngan, M., and Hanaoka, K. (2019a). Face recognition vendor test (frvt) part 2: Identification.
- Grother, P., Ngan, M., and Hanaoka, K. (2018). Ongoing face recognition vendor test (frvt) part 1: Verification. *National Institute of Standards and Technology*.
- Grother, P., Ngan, M., and Hanaoka, K. (2019b). *Face recognition vendor test (frvt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD.
- Guo, G. and Zhang, N. (2019). A survey on deep learning based face recognition. *Computer vision and image understanding*, 189:102805.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*.
- Jianhong, X. (2008). Kpca based on ls-svm for face recognition. In *2008 Second International Symposium on Intelligent Information Technology Application*, volume 2, pages 638–641. IEEE.
- Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114.
- Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863.
- Kim, M., Jain, A. K., and Liu, X. (2022). Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759.

- Kim, M., Liu, F., Jain, A., and Liu, X. (2023). Dcfac: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725.
- Kim, M., Sagong, M.-C., Nam, G. P., Cho, J., and Kim, I.-J. (2024). Vigface: Virtual identity generation model for face image synthesis. *arXiv preprint arXiv:2403.08277*.
- Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., and Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1931–1939.
- Kolf, J. N., Rieber, T., Elliesen, J., Boutros, F., Kuijper, A., and Damer, N. (2023). Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 806–816.
- Kong, R. and Zhang, B. (2011). A new face recognition method based on fast least squares support vector machine. *Physics Procedia*, 22:616–621.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220.
- Melzi, P., Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Lawatsch, D., Domin, F., and Schaubert, M. (2023). Gandiffac: Controllable generation of synthetic datasets for face recognition with realistic variations. *arXiv preprint arXiv:2305.19962*.
- Melzi, P., Tolosana, R., Vera-Rodriguez, R., Kim, M., Rathgeb, C., Liu, X., DeAndres-Tame, I., Morales, A., Fierrez, J., Ortega-Garcia, J., et al. (2024). Fracsyn challenge at wacv 2024: Face recognition challenge in the era of synthetic data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 892–901.
- Meng, Q., Zhao, S., Huang, Z., and Zhou, F. (2021). Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234.
- Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59.
- Oja, E. and Hyvarinen, A. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

- Papantoniou, F. P., Lattas, A., Moschoglou, S., Deng, J., Kainz, B., and Zafeiriou, S. (2024). Arc2face: A foundation model of human faces. *arXiv preprint arXiv:2403.11641*.
- Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.
- Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., and Tao, D. (2021). Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- Sengupta, S., Cheng, J., Castillo, C., Patel, V., Chellappa, R., and Jacobs, D. (2016). Frontal to profile face verification in the wild. In *IEEE Conference on Applications of Computer Vision*.
- Shahreza, H. O., Ecabert, C., George, A., Unnervik, A., Marcel, S., Di Domenico, N., Borghi, G., Maltoni, D., Boutros, F., Vogel, J., et al. (2024). Sdfr: Synthetic data for face recognition competition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE.
- Shahreza, H. O. and Marcel, S. (2024). Hyperface: Generating synthetic face recognition datasets by exploring face embedding hypersphere. *arXiv preprint arXiv:2411.08470*.
- Sharir, G., Noy, A., and Zelnik-Manor, L. (2021). An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.
- Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650.
- Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86.

- Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed.*, Cham: Springer International Publishing, 10(3152676):10–5555.
- Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274.
- Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A. K., Duncan, J. A., Allen, K., et al. (2017). Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98.
- Wolf, L., Hassner, T., and Taigman, Y. (2008). Descriptor based methods in the wild. In *Workshop on faces in'real-life'images: Detection, alignment, and recognition*.
- Wu, H., Singh, J., Tian, S., Zheng, L., and Bowyer, K. W. (2024). Vec2face: Scaling face dataset generation with loosely constrained vectors. *arXiv preprint arXiv:2409.02979*.
- Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. (2023). Demystifying clip data. *arXiv preprint arXiv:2309.16671*.
- Yeung, M., Teramoto, T., Wu, S., Fujiwara, T., Suzuki, K., and Kojima, T. (2024). Variface: Fair and diverse synthetic dataset generation for face recognition. *arXiv preprint arXiv:2412.06235*.
- Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., et al. (2021). Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502.